Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Zeeshan Ahmad (zeeshanahmad00789@gmail.com) **Contribution:** · Outlining project plan. • Data wrangling and maintaining data integrity. • EDA. • Data Preprocessing • Feature Engineering. • Model Building.
Please paste the GitHub Repo link.
GitHub Link: - https://github.com/Zeeshan00789/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING-

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The Netflix movies and TV shows clustering dataset consist of various categorical and numerical columns. The dataset consist of total of 12 columns such as 'show_id', 'type', 'title', 'director', 'added_year', 'added_month', 'added_day', 'target_ages' etc. I started by importing all the required libraries and dataset.

I followed step-by-step processes for the project like data collection, data cleaning, EDA, Visualization, Model Training and Testing.

In EDA I started by checking the head of the dataset, it contains various categorical and numerical columns. I dropped some categorical features which are irrelevant such as date_added. Further I check unique values and duplicate values in dataset. Later I count the duplicate values and drop them. I also worked on NAN values, I filled up all the NAN values with ‘unknown’. I created columns as year, month and date and started by univariate analysis. Later I do some visualizations such as boxplots to find outliers, count plot, barplot, and heatmap for inferences. I also explore the genres column to find different types of genres later I find top 15 genres with the help of countplotDocumentaries, standup comedy, Dramas and international movies are the top most genres in Netflix. I also plotted word cloud for description column. Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

Top 20 countries are also plotted with the help against ratings to find out insights. Here we can say that shows with TV-MA rating are highest in Belgium, Brazil, Italy, Turkey, Australia, Mexico, Germany, South korea, Spain, Japan, France, Canada, United Kingdom, United States. India, China, Egypt, Hongkong and Taiwan has highest number shows rated as TV-14. A countplot is also drawn to find wheather Netflix is focusing more on TV shows or movies. From 2017 the number of movies released is more than TV shows. Later we do text cleaning. Text cleaning is task-specific and one needs to have a strong idea about what they want their end result to be and even review the data to see what exactly they can achieve. We use the PCA to reduce the dimensions. I used k means clustering algorithm. The K-Elbow Visualizer implements the “elbow” method of selecting the optimal number of clusters for K-means clustering. The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center. We choose number of clusters=15. Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]